[BoldFont=Fira Sans SemiBold]Fira Sans Book Fira Mono

popquizbox[1] colback=nipun-lightblue!10, colframe=nipun-blue, boxrule=2pt, arc=3pt, left=8pt, right=8pt, top=8pt, bottom=8pt, title= **Quick Quiz 1**, fonttitle=**, coltitle=nipun-white, colbacktitle=nipun-blue, enhanced, attach boxed title to top left=xshift=0pt, yshift=-2pt, boxed title style=arc=3pt, boxrule=0pt**

definitionbox[1] colback=nipun-green!8, colframe=nipun-green, boxrule=1.5pt, arc=2pt, left=6pt, right=6pt, top=6pt, bottom=6pt, title= **Definition: 1**, fonttitle=**, coltitle=nipun-white, colbacktitle=nipun-green**

examplebox[1] colback=nipun-orange!8, colframe=nipun-orange, boxrule=1.5pt, arc=2pt, left=6pt, right=6pt, top=6pt, bottom=6pt, title= **Example: 1**, fonttitle=**, coltitle=nipun-white, colbacktitle=nipun-orange**

keypointsbox colback=nipun-blue!8, colframe=nipun-blue, boxrule=1.5pt, arc=2pt, left=6pt, right=6pt, top=6pt, bottom=6pt, title= **Key Points**, fonttitle=**, coltitle=nipun-white, colbacktitle=nipun-blue,**

Places where you will see unsupervised learning

- It can be used to segment the market based on customer preferences.

Places where you will see unsupervised learning

- It can be used to segment the market based on customer preferences.
- A data science team reduces the number of dimensions in a large dataset to simplify modeling and reduce file size.

Clustering

REQUIREMENTS: A predefined notion of similarity/dissimilarity.

REQUIREMENTS: A predefined notion of similarity/dissimilarity.

Examples:

Market Segmentation: Customers with similar preferences in the same groups. This would aid in targeted marketing.

$$WCV\left(C_i\right) = \frac{1}{|C_i|}_{PROTECTED_0}$$

$$\mathrm{WCV}(C_i) = \frac{1}{|C_i|} \sum_{a \in C_i} \sum_{b \in C_i} ||x_a - x_b||_2^2$$

where $|C_i|$ is the number of points in C_i

Then,

$$
\begin{aligned}
WCV\left(C_i\right) &= \frac{1}{|C_i|} \sum_{a \in C_i} \sum_{b \in C_i} ||x_a - x_b||_2^2 \\
&= 2 \sum_{a \in C_i} ||x_a - x_i||_2^2
\end{aligned}
$$

Then,

$$
\begin{aligned}
WCV\left(C_i\right) &= \frac{1}{|C_i|} \sum_{a \in C_i} \sum_{b \in C_i} ||x_a - x_b||_2^2 \\
&= 2 \sum_{a \in C_i} ||x_a - x_i||_2^2
\end{aligned}
$$

This shows that K-Means gives the **local minima**.

Hierarchal Clustering

There is no need to specify K at the start

There is no need to specify K at the start

k_bad_1.png	k_bad_2.png

Examples where K-Means fails

Algorithm for Hierarchal Clustering

1. Start with all points in a single cluster

Algorithm for Hierarchal Clustering

1. Start with all points in a single cluster

   2.1 Identify the 2 closest points

h_e_1.png

Algorithm for Hierarchal Clustering

1. Start with all points in a single cluster

   2.1 Identify the 2 closest points
   2.2 Merge them

h_e_1.png

Algorithm for Hierarchal Clustering

1. Start with all points in a single cluster
2. Repeat until all points are in a single cluster
   2.1 Identify the 2 closest points
   2.2 Merge them

h_e_1.png

h_e_2.png

Joining Clusters/Linkages

Joining Clusters/Linkages

Complete

Max inter-cluster similarity

Joining Clusters/Linkages

Complete

Max inter-cluster similarity

Single

Min inter-cluster similarity

Joining Clusters/Linkages

Complete
Max inter-cluster similarity

Single
Min inter-cluster similarity

Centroid
Dissimilarity between cluster centroids

More Code

Google Colab Link